Alternate Views of AGD

BIG DATA OPTIMIZATION - IE 598
HONGYI LI AND SAMANTHA THRUSH
29 NOV. 2016
Overview

- Introduction/Overview of AGD
- The Geometrical Interpretation (Bubek, Lee, and Singh, 2015)
- The Game Theory Perspective (Lan and Zhou, 2015)
- The ODE Viewpoint (Su, Boyd and Candes, 2015)
- Questions
What you know and love: works on L-smooth and $\mu$-strongly convex functions ($f$)

- L-smooth convergence rate: $O\left(\frac{LD^2}{t^2}\right)$

- L-smooth and $\mu$-strongly convex convergence: $O\left(\frac{(\sqrt{\kappa} - 1)^2 t}{\sqrt{\kappa}+1}\right)$

- Created in 1983 by Yurii Nesterov for smooth functions, has been generalized to non-smooth or composite problems

- Pure “algebraic trick”; no true physical intuition behind original method.

- Optimal convergence rates
AGD Scheme

- A two step process: allow $x_0 = y_0$ for $t = 0, 1, 2, 3, \ldots$ and:
  $$\alpha_{t+1}^2 = (1 - \alpha_{t+1})\alpha_t^2 + \frac{c}{L}\alpha_{t+1}$$

- Allow $x$ and $y$ to be updated in the following fashion:
  $$x_{t+1} = y_t - \frac{1}{L} \nabla f(y_t)$$
  $$y_{t+1} = x_{t+1} + \beta_t (x_{t+1} - x_t)$$

Where the following is true:
  $$\beta_t = \frac{\alpha_t (1 - \alpha_t)}{\alpha_t^2 + \alpha_{t+1}}$$

For the ‘83 Nesterov scheme, the following was true instead:
  $$\beta_t = \frac{\sqrt{L} - \sqrt{c}}{\sqrt{L} + \sqrt{c}}$$
The Geometrical Interpretation
The Geometrical Interpretation

From smoothness and strongly convexity, we have:

\[ x_+ = x - \frac{1}{L}\nabla f(x) \]

\[ x_{++} = x - \frac{1}{\mu}\nabla f(x) \]

\[ \|x_* - x_{++}\|^2 \leq \left( 1 - \frac{1}{\kappa}\right) \frac{\|
abla f(x)\|^2}{\mu^2} - \frac{2}{\mu} (f(x_+) - f_*) \]
The Geometrical Interpretation

\[ R^2 = \left( 1 - \frac{1}{\kappa} \right) \frac{||\nabla f(x)||^2}{\mu^2} - \frac{2}{\mu} (f(x_+) - f_0), \quad \kappa = \frac{L}{\mu} \]

It looks like this:
The Geometrical Interpretation

\[ B(0,1) \cap B(g, (1 - \varepsilon)\|g\|^2) \subseteq B(x, 1 - \varepsilon) \]
The Geometrical Interpretation

\[ x_\ast \in B(x_0, R_0^2) \]
\[ x_\ast \in B(x_{0++}, (1 - \kappa) \frac{\|\nabla f(x_0)\|^2}{\mu^2}) \]

Let \( x_1 \) be the \( x \) in previous graph, then repeat:

\[ \|x_k - x_\ast\|^2 \leq \left(1 - \frac{1}{k}\right)^k R_0^2 \]

weaker than \( \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \) converge
The Geometrical Interpretation

\[ B(0, 1 - \varepsilon g^2 - \delta) \cap B(a, (1 - \varepsilon)g^2 - \delta) \subseteq B(x, 1 - \sqrt{\varepsilon} - \delta), \|a\| \geq g \]
The Geometrical Interpretation

Intuition: \[ \|x_+ - x^++\|^2 \leq \left(1 - \frac{1}{\kappa}\right)\frac{\|\nabla f(x)\|^2}{\mu^2} - \frac{2}{\mu} (f(x_+) - f_*) \], set \( \delta = \frac{2}{\mu} (f(x_+) - f_*) \), can achieve \( \sqrt{\kappa} \).
The Geometrical Interpretation

\[ x_{k+1} = \text{linear search} \ (c_k, x_k) \]

\[ x_* \in B(c_k, R_k^2 - \frac{\| \nabla f(x_{k+1}) \|^2}{\mu^2 \kappa} - \delta) \]

\[ x_* \in B(x_{k+1} + (1 - \frac{1}{\kappa}) \frac{\| \nabla f(x_{k+1}) \|^2}{\mu^2} - \delta) \]

Let \( c_k, R_k^2 \) be updated knowing \( x_* \in B(c_k, R_k^2) \):

\[ R_k^2 \leq \left( 1 - \frac{1}{\sqrt{\kappa}} \right)^k R_0^2 \]

\[ \lesssim O((\frac{\kappa-1}{\kappa+1})^{2k}) \text{ when } \kappa \text{ large, weaker than } (\frac{\sqrt{\kappa-1}}{\sqrt{\kappa+1}})^{2k} \text{ converge} \]
The Geometrical Interpretation

Fact of big and small balls:
The Geometrical Interpretation

\[ x_+ = \text{linear search} \left( x, x - \nabla f(x) \right) \]

\[ x_* \in B(c_k, R_k^2 - \frac{2}{\mu} (f(x_k^*) - f(x_{k+1}))) \]

\[ x_* \in B(x_{k+1}^+, \frac{\|\nabla f(x_{k+1})\|^2}{\mu^2} - \frac{2}{\mu} (f(x_{k+1}) - f(x_{k+1}))) \]

At least as fast as previous one

All method need (lower bound) \( \mu \)!
The Geometrical Interpretation

Example: Logistic Regression (L2)
The Game Theory Perspective
The Game Theory Perspective (GTP)

- Two different “Games” and two different techniques.
- A different subset of Gradient Descent
- Accelerated Gradient Descent is a special case of the Primal-Dual gradient method
GTP - Inspiration

- Started off with a scheme that was not unlike a buyer/seller scheme.
- Had two updated variables: primal and dual, and was in search of optimal solution of a saddle point problem.
- Needed both to converge to the same value.
- Two different quantities being updated that are dependent on each other, not unlike a two player game.
- Buyer/seller scheme can be thought of in game-theory terms: buyer and seller both want to “win” (get/keep the most money), so they look for the best order quantity and product price.
Two players: Buyer (primal) and Seller (dual)

The goal of the game is to come to an equilibrium of cost and price (neither want to be taken advantage of, but both want a fair price/profit)

What they know:

- their local cost (buyer: $h(x) + \mu \omega(x)$, seller: $J_f(g)$)
- interactive cost (but not their partner’s)(buyer: revenue, supplier: $\langle x, g \rangle$)

Three different steps, taken iteratively.
Algorithm 1 The primal-dual gradient method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $g^0 = \nabla f(x^0)$.

for $t = 1, \ldots, k$ do

Update $(x^t, g^t)$ according to

$$\hat{x}^t = \alpha_t (x^{t-1} - x^{t-2}) + x^{t-1}. \tag{2.8}$$
$$g^t = \mathcal{M}_g (-\hat{x}^t, g^{t-1}, \tau_t). \tag{2.9}$$
$$x^t = \mathcal{M}_X (g^t, x^{t-1}, \eta_t). \tag{2.10}$$

end for
Saddle point problem

Seller predicts demand $\bar{x}_t$ by looking two steps into the past:

Seller determines price to best bring home the Benjamins (essentially, we are maximizing $< \bar{x}^t, g > - J_f(g)$, regularized by: $D_f(g^{t-1}, g)$ (the dual prox-function) with weight $\tau_t \geq 0$.

Buyer then tries to find the best number of objects to buy in order to minimize the cost. Essentially minimizing $h(x) + \mu \omega(x) + < x, g >$, regularized by the primal prox-function $P(x^{t-1}, x)$ with weight $\eta_t \geq 0$

Will keep doing this iteratively until equilibrium is found (best order quantity and product price)

Price too high, won’t order much; price too low, orders will be huge and the seller gets taken advantage of.
GTP – Game 2 (Randomized Prime Dual Gradient)

- More than 2 players: 1 buyer and m number of suppliers
- The goal of the game is to come to an equilibrium of cost and price (neither want to be taken advantage of, but both want a fair price/profit)
- What they know:
  - their local cost (buyer: \( h(x) + \mu \omega(x) \), sellers: \( J_f(g) \))
  - interactive cost (but not their partner’s) (buyer: revenue, supplier: \( \langle x, g \rangle \))
- But! The buyer must purchase the same amount of products from each supplier
- And, a random supplier can make a price change during each iteration according to the predicted demand
Four different steps, taken iteratively.
Solves saddle point problem (again)
Randomization allows you to reduce the total number of gradient evaluations needed (but more primal prox-mapping)
Must make sure that f’s are differentiable over $\mathbb{R}^n$ (or else method might fail)
Algorithm 3 A randomized primal-dual gradient (RPDG) method

Let $x^0 = x^{-1} \in X$, and the nonnegative parameters $\{\tau_i\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ be given.
Set $y_i^0 = \nabla f_i(x^0)$, $i = 1, \ldots, m$.

for $t = 1, \ldots, k$ do

Choose $i_t$ according to $\text{Prob}\{i_t = i\} = p_i$, $i = 1, \ldots, m$.
Update $z^t = (x^t, y^t)$ according to

$$
\tilde{x}^t = \alpha_t(x^{t-1} - x^{t-2}) + x^{t-1}.
$$

$$
y_i^t = \begin{cases} 
M_{Y_i}(-\tilde{x}^t, y_{i-1}^t, \tau_i), & i = i_t, \\
y_{i-1}^t, & i \neq i_t.
\end{cases}
$$

$$
\tilde{y}_i^t = \begin{cases} 
p_i^{-1}(y_i^t - y_{i-1}^t) + y_{i-1}^t, & i = i_t, \\
y_{i-1}^t, & i \neq i_t.
\end{cases}
$$

$$
x^t = M_X(\sum_{i=1}^m \tilde{y}_i^t, x^{t-1}, \eta_t).
$$

end for
Sellers predict demand $\bar{x}^t$ by looking two steps into the past.

All sellers determine price to best bring home the Benjamins.

Then one random seller changes their price additionally, so that they have a small advantage (price-wise).

Buyer then tries to find the best number of products to buy from all sellers in each iteration so as to minimize cost.

Will keep doing this iteratively until equilibrium is found. (Ideal cost, and ideal amount to buy)
Differences between AGD and this method

- AGD is a special case of the primal-dual gradient methods presented here:

  “... the computation of the gradient at the extrapolation point of the accelerated gradient method is equivalent to a primal prediction step combined with a dual ascent step (employed with the aforementioned dual prox-function) in the PDG method.”- Lan and Zhou

- Essentially, the primal-dual gradient method presented can be thought of as just re-labeling some constants found in AGD, but updates variables in the same manner in both methods.

- Can be considered as a stand in for AGD if you consider the more generalized algorithms (Not Nesterov’s originals)

- Convergence rates comparable to original AGD for first method
The ODE Viewpoint
The ODE Viewpoint – Inspiration & Precedent

- Traditional link between ODEs and optimization (Helmke and Moore 1996, and others). Examples
  - Linear regression via linearized Bregman iteration algorithm (Osher 2014)
  - Control design via a continuous-time Nesterov-like alg. (Durr and Ebenbauer 2012)
- Others
- Simply done (numerically) by taking extremely small steps in a numerical optimization scheme to allow method to converge to an ODE modeled curve.
Nesterov's accelerated gradient decent scheme:

\[
\begin{align*}
    x_{k+1} &= y_k - s\nabla f(y_k) \\
    y_{k+1} &= x_{k+1} + \frac{k}{k+3} (x_{k+1} - x_k)
\end{align*}
\]

What if \( s \to 0? \)

Let's \( \Delta t = \sqrt{s}, \) \( t_k = k\Delta t \)
The ODE Viewpoint

\[ X'' + \frac{3}{t} X' + \nabla f(X) = 0 \]

\[ y_{k+1} = x_{k+1} + \frac{k}{k + \gamma} (x_{k+1} - x_k) \rightarrow X'' + \frac{\gamma}{t} X' + \nabla f(X) = 0, \gamma \geq 3 \]
The ODE Viewpoint

- **Time Invariance:**
  \[ t \rightarrow ct, Y(\tau) = X(c\tau), Y'' + \frac{\gamma}{\tau} Y' + \frac{\nabla f(Y)}{c^2} = 0 \]

- **Rotation Invariance:**
  If \( X(t) \) minimizes \( f \), \( Q \) is rotation, \( g(x) = f(Qx) \), then \( QX(t) \) minimizes \( g \).

- **Initial Asymptotic:**
  When \( t \) small, \( X(t) \approx -\frac{\nabla f(x_0)t^2}{8} + x_0 + o(t^2) \)
The ODE Viewpoint

\( f \) \( L \)-smooth:

\[
\begin{align*}
  f(X(t)) - f_* & \leq \frac{2\|x_0 - x_*\|^2}{t^2} \\
  \int_0^\infty t(f(X(t)) - f_*) dt & < \infty
\end{align*}
\]

- Energy Function \( \varepsilon(t) \):

\[
\begin{align*}
  \frac{2t^2}{\gamma - 1} (f(X(t)) - f_*) + (\gamma - 1) \left\| X(t) - \frac{t}{\gamma - 1} X'(t) - x_* \right\|^2
\end{align*}
\]
The ODE Viewpoint

Proximal Gradient for $f = h + g$, $g$ $L$-smooth, $h$ non-smooth.
No ODE for $h$ not smooth.

$$f(x_k) - f_* \leq C \frac{\|x_0 - x_*\|^2}{(k + \gamma - 2)^2}, C \text{ depends on } \gamma > 3, s < 1/L.$$ 

$$\sum_{k=1}^{\infty} (k + \gamma - 1)(f(x_k) - f_*) < \infty$$

- **Discrete Energy Function** $\epsilon(k)$:

  $$\frac{2(k + \gamma - 2)^2}{\gamma - 1} s (f(x_k) - f_*) + (\gamma - 1) \|z_k - x_*\|^2$$

  $$z_k = \frac{k + \gamma - 1}{\gamma - 1} y_k - \frac{k}{\gamma - 1} x_k$$
The ODE Viewpoint

\[ f \text{ } L\text{-smooth, } \mu\text{-strongly convex: } (\gamma \geq 4.5) \]

\[ f(x(t)) - f_* \leq C \frac{\|x_0 - x_*\|^2}{t^3}, \ C \text{ depends on } \gamma, \mu. \]

- Energy Function \( \epsilon(t) \):

\[ t^3(f(X(t)) - f_*) + \frac{(2\gamma - 3)^2 t}{8} \left\| X(t) - \frac{2t}{2\gamma - 3} X'(t) - x_* \right\|^2 \]
The ODE Viewpoint

$\gamma = 3$, even for quadratic function, cannot be better than $O\left(\frac{1}{t^3}\right)$

$(\Delta t = \sqrt{s}, s = 0.000001 \text{ observe every 1000 steps})$
The ODE Viewpoint

- Old Guess: $f$ $L$-smooth, $\mu$-strongly convex:
  \[ f(X(t)) - f_* \sim O\left(\frac{1}{t^\gamma}\right) \]

- New Discovery (Journal of Machine Learning Research, published 9/16)
  - $f$ $L$-smooth, $\mu$-strongly convex:
    \[ f(X(t)) - f_* \leq O\left(\frac{1}{t^{3\gamma}}\right) \]
  - $f = h + g$, $g$ $L$-smooth, $\mu$-strongly convex:
    \[ f(X(t)) - f_* \leq O\left(\frac{1}{t^3}\right) \]
The ODE Viewpoint

Restart when $\frac{d\|X(t)\|^2}{dt} \leq 0$: keep $X(t)$ moving fast

$$y_{k+1} = x_{k+1} + \frac{k}{k+3} (x_{k+1} - x_k) \quad \xrightarrow{k=0} \quad y_{k+1} = x_{k+1}$$

$f$ $L$-smooth, $\mu$-strongly convex: ($\gamma = 3$)

$$f(X(t)) - f_* \leq C\|x_0 - x_*\|^2 \exp(-C't), \quad C, C' \geq 0 \text{ depend on } \mu, L.$$
The ODE Viewpoint

Example: LASSO ($f = h + g$)
Questions??